STATISTICAL METHODS FOR AUTOMATIC BRAIN SEGMENTATION

by

James K. Pringle

A thesis submitted to Johns Hopkins University in conformity with
the requirements for the degree of Master of Science

Baltimore, MD

December 2014

1 Abstract

Background: Automatic segmentation of the brain into cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM) has been of interest for over twenty years. As magnetic resonance imaging (MRI) has improved and new sequences have been developed, more and more data can be utilized to improve and accelerate segmentation algorithms.

Objective: To segment the brain into CSF, GM, and WM using multichannel MRI data (T1, T2, PD, FLAIR, water image, and MTC) with a multinomial logistic regression (MLR) model and to compare the results to software segmentations from FSL, FreeSurfer, and TOADS-CRUISE.

Methods: Within each subject, the different MRI sequences are co-registered to the water image within subject. Bias field correction and intensity normalization is then applied. The aligned T1 images are used as inputs for existing automatic segmentation software. FreeSurfer and TOADS anatomical segmentations are combined into CSF, GM, and WM. Our method uses MLR applied to normalized brain images. Models are further refined by adding spline terms to model possible non-linear associations.

Results: Measures of similarity—the Jaccard index, the dice index, and the confusion matrix—are presented to compare the results of existing software with those obtained from the new MLR method. Segmentations are also compared and rated by a radiology resident.

Conclusions: Results based on MLR are comparable to software segmentations. In some areas, they outperform existing software.

Readers: Dr. Ciprian Crainiceanu (advisor) and Dr. Ani Eloyan

2 Acknowledgements

I would like to thank the NIH and the AFNI group for teaching me and answering my numerous questions about MRI imaging. My advisor, Dr. Ciprian Crainiceanu, and my thesis reader, Dr. Ani Eloyan, have been immensely supportive and helpful during this writing process. Finally, I owe deep gratitude to my wife and her scientific know-how.

3 Table of Contents

4 List of Figures

5 List of Tables

6 Introduction

Magnetic Resonance Imaging (MRI) has come to be an indispensable tool for diagnosing and monitoring neurological diseases and for examining soft tissue in the body. MRI is non-invasive and provides imaging with good contrast in places like the brain (Rivest-Henault and Cheriet 2011).

One important operation with MRI is to label regions of interest, whether they be a multiple sclerosis lesion, certain anatomical features, the cerebral cortex, the whole brain, or something else (Lladó 2012; Fischl 2002; Hutton 2008; Smith 2002). Labeling the brain into cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM) has been of interest for more than twenty years, and many methods exist for brain segmentation into these three tissues (Bonar et al. 1993; Kikinis et al. 1992; West et al. 2012). A common theme in the literature is to compare these methods against gold standards: manually drawn reference segmentations verified by radiologists.

Only recently has software for brain segmentation become convincing (Rivest-Henault and Cheriet 2011). Withey and Koles identify three generations of segmentation software that can be categorized as unsupervised learning algorithms or automatic segmentation. The first generation uses simple thresholding and volume growing methods, the second generation incorporates uncertainty models and optimization, and the third generation is characterized by knowledge—typically in the form of atlases. Over time, software segmentations have agreed better and better with gold standards (Rivest-Henault and Cheriet 2011).

The advancement of MRI technology, along with its increasingly finer resolution and new MRI sequences (Cotten et al. 2009), has allowed more data to be acquired. Because of

this, for example, the errors due to the partial volume effect have become smaller (Gonzalez Ballester et al. 2000). Furthermore, multichannel data—data from more than one MRI sequence of the same subject—offers a richer feature space on which to perform analyses (Gordillo et al. 2013). This provides fertile opportunity for model and algorithm development and validation.

One of the major drawbacks of these algorithms is the large computation time needed to produce results. In one recent study, ANTs and FreeSurfer took on average 15.7 hours and 14.1 hours, respectively, to run (Tustison et al. 2014). Software tools that provide similar results more quickly would be valuable.

In this study, a supervised classifier without an atlas—multinomial logistic regression—is proposed. We focus on fast computational approaches with results comparable with current segmentation software. After a common preprocessing pipeline, the MLR model performs well on the subjects available for the study. The method, results, and their comparison with current segmentation software are presented below.

7 Methods

In this section we introduce two classification algorithms: the multinomial logistic regression (MLR) statistical model and the *k*-Nearest Neighbors classifier (*k*-NN). MLR is the focus of this study and *k*-NN is used as a reference. Before either algorithm is run, all MRI sequences are co-registered within subject to the water image, and the images are normalized locally. The segmentation results of FSL, FreeSurfer, and TOADS-CRUISE are used as training sets separately for the two algorithms. These three programs create their own skull-stripped brain masks. The results are cross-validated by testing on one subject and training on the others.

Furthermore, measures of similarity—the Jaccard index, the dice index, and the confusion matrix—are derived to compare the output of the software to the two models. Finally, the results are compared to the output of the software on slices of interest.

7.1 Study population

There are four subjects available. All four subjects are healthy adults that are controls in an ongoing study of multiple sclerosis. The water image is not widely utilized; thus segmentation results using the water image may not be applicable to other studies.

7.2 Image acquisition

The battery of images—four T1 echoes, T2, proton density (PD), fluid-attenuated inversion recovery (FLAIR), water image, and three magnetization transfer contrast (MTC) volumes—was acquired for all subjects on the same 3T MRI scanner. The T1 was calculated as the quadratic mean of the four T1 echoes. See Table 1 for a summary of the protocol

parameters and resolution for each MRI sequence. The original images were used as a starting point for preprocessing.

7.3 Image preprocessing

Before running the classification algorithms for segmentation, the images were preprocessed with the AFNI software. Software segmentations were obtained using the FSL, FreeSurfer, TOADS-CRUISE, and MIPAV software packages.

Preprocessing has many stages, and each image must pass through each phase before a proper and reliable analysis can occur. Figure 1 outlines the preprocessing steps and generation of software segmentations. Since the proposed analysis uses intensities from different MRI sequences, co-registration, or alignment, is the first step. This ensures that the same voxel index across the different MRI sequences refers to the same location in space relative to the subject's head. The co-registered images are used as inputs for the segmentation software. After that, the bias field is removed so that image intensities are more uniform within tissue classes for the same subject. Finally, images are centered and scaled with respect to different brain masks, making units comparable across MRI sequences. This last step is necessary for running the *k*-NN algorithm. The resulting images obtained after preprocessing are aligned and corrected for spatial inhomogeneity. These images can be used for further statistical modeling and associated software.

As an aid to explain the pipeline, Figure 2 shows a representative axial slice of a PD image at each stage of preprocessing. Each stage of preprocessing is explained in further detail below.

7.3.1 Alignment

Due to a variety of reasons, such as inter-subject anatomical variability, time between consecutive MRI sequences in a protocol, human movement, and scanner variability, brains in different MRI scans do not overlap perfectly. For experiments involving multiple subjects, it is common to register all images for all subjects to a single standard space, such as the Talairach or MNI space. A problem with such approaches is that the associated transformation may lead to distortions of the brain that may not be anatomically correct or may induce artifacts. However, the two classification algorithms in this study do not require such registration across subjects. Consequently, within each subject, all MRI images were co-registered, aligning each image to the water image via affine transformations. Reference Figure 2A – 2C.

Based on visual inspection, the water image used as the base image in the affine transformations gave the best results. There were exceptions, and difficulties sometimes arose when registering PD and T2 images. Using a multi-step alignment seemed to address this problem. Since the other images (T1, FLAIR, and MTC) had already passed through this section of the pipeline, it then became possible to align to these other images. This workaround with "helper data" led to successful alignment of all images.

7.3.2 Bias field removal

After registration, the next step was to remove the bias field. The bias field is spatial inhomogeneity induced by the proximity of the radiofrequency (RF) coils in the MRI scanner. Correcting for spatial inhomogeneity is essential for analysis based on image intensities because it enables comparison between different regions in the brain that contain the same tissue class. For example, if white matter in the lower and upper part of the brain

corresponds to different image intensities, then most algorithms would fail to recognize both areas as white matter.

Let the observed MRI intensity at voxel (location) i during scan sequence z be modeled as

$$y_z(i) = g_z f_z(i) v_z(i) + \varepsilon(i).$$

An underlying image v is assumed to be dependent only the specific MRI sequence and tissue classes (Vovk, et al. 2011). Distortions from the underlying image come from signal gain g, from the spatially smooth and spatially dependent bias field f, and from noise ε. In this model, getting a good estimate of $kg_z f_z(i)$, where k is a scalar value, yields a scaled version of v.

In most MRI sequences, one of CSF, GM, or WM is the most hyperintense tissue (has the highest intensity). For example, in a T1 image, WM is the brightest. Since these tissue classes are ubiquitous in the brain, it is possible to trace the bias field by following the brightest tissue class throughout the brain. Under this "brightness ubiquity" assumption and the assumption of smoothness in the bias field, it is possible to interpolate the bias field at other voxels other than the brightest tissue class. To that end, the following algorithm is devised to calculate $kg_z f_z^*(i)$, where $f_z^*(i)$ is an estimate of the bias field.

Let $R_{z,45}(i)$ be the collection of all intensities at voxels within 45 mm of voxel i in scan z. Let $p_{90}(R_{z,45}(i))$ be the 90th percentile of $R_{z,45}(i)$. Then $p_{90}(R_{z,45}(i))$ shows the intensity trend of the brightest tissue class throughout the brain. Note that since $p_{90}(R_{z,45}(i))$ is based on image intensities, it includes information about the scan gain. Based on the assumptions above, $p_{90}(R_{z,45}(i))$ is a good estimate of $kg_z f_z^*(i)$.

There are two parameters in the bias field estimation algorithm presented above: the radius of the neighborhood and the percentile of the neighborhood. If the radius of the neighborhood is too small, it is possible that $R_{z,45}(i)$ may not capture enough bright intensities near voxel i. If the radius is too large, then $R_{z,45}(i)$ and $R_{z,45}(i')$ for $i \neq i'$ may become too similar to have meaningful distinctions in the percentile function. In other words, the bias field could be too smooth. A radius of 45 mm is chosen because it gives good empirical results, and the percentile is chosen so that it captures a characteristic intensity of the brightest tissue.

New images are generated from the bias field estimation. For each scan sequence z let

$$y_z^*(i) = \frac{y_z(i)}{k g_z f_z^*(i)} \approx k^{-1} v_z(i) + \varepsilon^*(i).$$

Reference Figure 2D and 2E. Because of the assumption that the 90th percentile in a local neighborhood contains information about scan gain, the gain parameter cancels out in the division. Hence, scans of the same sequence on different subjects are brought to a common scale. This means that in addition to correcting for spatial inhomogeneity, the algorithm normalizes the data. See Figure 1 for a comparison of densities of intensities across subjects. There are other very successful bias correction strategies including N3 and N4 (Tustison, et al. 2010). For the purposes of this research we will use the local filtering approach described above.

7.3.3 Centering and scaling

Because *k*-NN uses Euclidean distance to decide the nearest neighbors, the data need to be centered and scaled. Even though the bias field removal does a normalization for

the same image across subjects, it does not necessarily center and scale the data across images. For a particular brain mask M, let the normalized data be

$$n_z^M(i) = \frac{y_z^*(i) - \mu_z^M}{\sigma_z^M}$$

where μ_z^M is the mean intensity and σ_z^M is the standard deviation of $y_z^*(i)$ as i ranges over the voxels contained in M. Hence the data are normalized image by image with respect to the specific brain mask M. The specific masks used are the output of existing software segmentations. This technique was proposed in Shinohara, et al., 2011, and a more refined version was proposed in Shinohara, et al., 2014.

7.3.4 Generation of software segmentations

Three sets of software gold standards were generated. In general, there are two phases: brain extraction (also known as skull stripping) and segmentation. The software and their specific tools are the following: TOADS-CRUISE with SPECTRE for brain extraction and TOADS for brain segmentation; FSL with BET for brain extraction and FAST for brain segmentation; and FreeSurfer with the `recon-all` command.

This software is not fail-proof, and certain workarounds were required. The most common problem was a grossly incorrect brain mask. The software can be too greedy, and the brain mask can include significant swaths of extraneous voxels such as, for example, parts of the skull and skin surrounding the brain. At other times, the software may miss large parts of the brain. In these cases, the helper data principle applies. Since the automated programs are broken into steps—extracting the brain and then segmenting the brain—it is possible to use skull-stripped brains from other software platforms that performed better on a particular brain image. In general, the brain extraction step in all three programs is not

perfect. However, relatively few false positives and negatives are tolerable when the alternative is obtaining a manually drawn gold standard.

Other problems may arise from the software as well. For example, TOADS does not carry over file header information. TOADS may also reorient the data (orientation refers to the transformation from the 1D array in which the data is stored to the 3D array viewed on screen). FreeSurfer may change the origin information in the header, change the image dimensions, or leave its segmentation in the template space it uses. These may necessitate alignment to the original T1 image and resampling, and care must be taken to use nearest neighbor interpolation (not linear or cubic, etc.).

While FSL calculates segmentations into CSF, GM, and WM, both FreeSurfer and TOADS generate richer anatomical segmentations. FreeSurfer labels 45 different regions labeled, while TOADS labels 10 different anatomical features. The output of these two programs can be condensed to CSF, GM, and WM by marking each anatomical feature as one of the three tissue classes. With TOADS, the labeling is straightforward. The reclassification for FreeSurfer is based on Kowkabzadeh's work (Kowkabzadeh 2010).

7.4 Classification and statistical modeling

This study compares two new classification algorithms developed for a richer set of brain imaging predictors to existing segmentation software. Both prediction algorithms were trained and tested using exhaustive cross-validation. First, a classification set is chosen based on existing software segmentations. More precisely, segmentations were obtained from each existing segmentation software and each was used in turn as a gold standard for the new methods. Then, in one round of cross-validation, one subject is labeled as the test data set. The other three subjects are used to train the models and make predictions for the test data

set. The process is repeated for the four subjects for a total of four rounds of cross-validation. Once predictions are available for all four subjects, the similarity between the predictions and the original classifications is evaluated. Finally, the whole process is repeated for the different software segmentations.

One limitation of the approach is that for this part of the analysis we did not have true gold standard segmentations. Instead, each software was used to provide a gold standard and the comparison was done with the new approaches based on this gold standard. Thus, the newly proposed statistical methods are tuned to each software segmentation separately. To circumvent some of these problems, it may be worth considering fusing the prediction results using a weighting scheme. Here we have decided to use this approach and a human observer who compared the results of the new algorithms with the ones of pre-existing software.

The *k*-NN classifier and the MLR statistical model were implemented in the R environment (version 3.1.1, R Foundation for Statistical Computing) using the packages `FNN` and `nnet`.

7.4.1 *k*-Nearest Neighbors

The *k*-NN classifier is a supervised learning algorithm (Devroye, Györfi and Lugosi 1996). In basic terms, for a new data point with a set of descriptors (covariates), the algorithm takes the nearest k data points (measured by Euclidean distance between the sets of descriptors) from the training data set. The most common classification among the k data points is assigned to the new data point (ties are broken randomly). According to Devroye, Györfi, and Lugosi, 1996, a rule of thumb for choosing k is to take the square root of the size of the training set.

Only a subset of the battery of images was used for *k*-NN. Since T1 is a function of the four T1 echoes, the T1 was kept and the four echoes were dropped. Finally, since the three MTC images are similar and the first two MTC images have much lower contrast in the brain than the third, the third MTC image was kept and the others were dropped.

The training data set can be quite large—on the order of five million rows. This makes running the *k*-NN algorithm computationally impossible. Therefore, a random subset of 5% of the rows of the training set was used to classify the test set, and this was repeated twenty times. To get a final result, the most common classification for each voxel was selected with ties being broken at random.

7.4.2 Multinomial Logistic Regression Statistical Model

Multinomial logistic regression (MLR) is a supervised prediction model for the case when the number of classes to be predicted is larger than two. MLR is a direct generalization of logistic regression, which is used for the case when one is interested in prediction of only two classes (e.g. dead/alive or brain/non-brain). We provide a brief introduction to MLR.

Let $\boldsymbol{x}(i)$ represent the vector of covariates/descriptors for voxel i. Let $Y(i)$ take values in $\{1, 2, \cdots, J\}$, which represent the J tissue classes that we are trying to predict. In our case $J = 3$ (CSF, grey matter, and white matter). The method can be easily generalized to any number of classes. The model is designed for $\pi_j\{\boldsymbol{x}(i)\}$, the probability of voxel i with covariates $\boldsymbol{x}(i)$ belonging to tissue class j. We denote this in compact form as

$$\pi_j\{\boldsymbol{x}(i)\} = P\{Y(i) = j \mid \boldsymbol{x}(i)\}.$$

Then the multinomial logistic regression model is

$$\log\frac{\pi_j\{\boldsymbol{x}(i)\}}{\pi_J\{\boldsymbol{x}(i)\}} = \alpha_j + \boldsymbol{\beta}'_j\boldsymbol{x}(i)$$

where J is a reference tissue class. The interpretation of α_j is the log odds ratio of belonging to category j as compared to the reference category, given that all covariates are zero. The interpretation of $\boldsymbol{\beta_j}$ is the change in the same log odds ratio associated with a unit increase in $\boldsymbol{x}(i)$. Since $\sum_{j=1}^{J} \pi_j\{\boldsymbol{x}(i)\} = 1$, it is straightforward to calculate that

$$\pi_j\{\boldsymbol{x}(i)\} = \frac{\exp\{\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}(i)\}}{1 + \sum_{h=1}^{J-1} \exp\{\alpha_h + \boldsymbol{\beta}_h'\boldsymbol{x}(i)\}} \quad (1)$$

with $\alpha_J = 0$ and $\boldsymbol{\beta}_J' = \mathbf{0}$. The log likelihood of the MLR model is

$$\log \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \pi_j\{\boldsymbol{x}(i)\}^{y_{ij}} \right],$$

where y_{ij} is 1 if $Y(i) = j$ and 0 otherwise. After substituting in equation (1) into the log likelihood, MLR uses the Newton-Raphson optimization algorithm to maximize the log likelihood. See Agresti 2002 for a more in-depth treatment of this topic.

Thus many models can be obtained simply by adding MRI sequences or functions of those sequences evaluated at the voxel level. We start by describing our basic model in which all sequences are used. While local information around the voxel could be used, here we focus only on the intensities at the particular voxel. Thus for this model,

$$\boldsymbol{x}(i) = \{y_{T1}^*(i), y_{T2}^*(i), \cdots, y_{MTC\,3}^*(i)\}$$

the set of normalized image intensities. In other words, the MLR model is

$$\begin{aligned} \log \frac{\pi_j\{\boldsymbol{x}(i)\}}{\pi_J\{\boldsymbol{x}(i)\}} = {} & \alpha_j + \beta_{j,1} y_{T1}^*(i) + \beta_{j,2} y_{T2}^*(i) + \beta_{j,3} y_{PD}^*(i) + \beta_{j,4} y_{FL}^*(i) + \beta_{j,5} y_{WT}^*(i) \\ & + \beta_{j,6} y_{T1E1}^*(i) + \beta_{j,7} y_{T1E2}^*(i) + \beta_{j,8} y_{T1E3}^*(i) + \beta_{j,9} y_{T1E4}^*(i) \\ & + \beta_{j,10} y_{MTC1}^*(i) + \beta_{j,11} y_{MTC2}^*(i) + \beta_{j,12} y_{MTC3}^*(i) \end{aligned}$$

The covariates in this model are the twelve bias field corrected images. For example, $y_{FL}^{*}(i)$ is the FLAIR normalized intensity.

The model is trained on the data from three subjects, and estimates of the parameters α and $\boldsymbol{\beta}$ are obtained. These are then plugged in to calculate, for each voxel i and tissue type j, the probability $\pi_j\{\boldsymbol{x}(i)\}$. The tissue class with the largest probability is assigned to that voxel. The MLR model was inspired by and is closely related to the OASIS and SUBLIME approaches that were designed for multiple sclerosis lesion segmentation (Sweeney, et al. 2013; Sweeney, et al. 2013). Both of these approaches use a logistic regression and multiple image sequences to build a classifier.

7.4.3 MLR model refinement

The first MLR model is a basic regression, and it can be further refined and improved. First, for the same reasons as in *k*-NN, the four T1 echoes and the first two MTC images are dropped. Second, the association between the log odds ratios of the voxel being in a particular tissue class and voxel level image intensities may not be linear. Thus, we can actually add regression splines that can capture possible nonlinear associations. To find the knots for the linear splines, the mode of the density of each tissue mask is determined image by image in the training set.

7.5 Validation of results

A variety of methods are used to compare the results from existing software segmentation algorithms with the results from the new proposed methods. Results are compared using the Jaccard index and the dice index for each subject and tissue class. The Jaccard index is a measure of similarity between two sets of voxels A and B defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

where $|A|$ is the number of voxels in set A. The dice index is another measure of similarity, and it is defined as

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|}.$$

Both indices indicate perfect overlap or no intersection when they are equal to 1 or 0, respectively. An overall Jaccard index, dice index, and confusion matrix is calculated and reported for each pair of classifier and software segmentation.

The results were also examined visually. Four pre-determined slices of interest per subject—two axial slices, one sagittal, and one coronal—were presented and the diverse segmentations were compared and scored by a radiology resident. Each software segmentation was compared to results from the two new proposed segmentation methods on these pre-determined slices. All pairs of images were randomized and the order within each pair was randomized. As a reference, the same slices from both the T1 and the difference image were displayed alongside the two segmentations. An example is presented in Figure 4. All images were rated as an integer from -2 to 2. A positive number refers to segmentations from the newly proposed methods, and a negative number refers to software segmentations. An absolute value of "2" indicates that the image is a much better segmentation than the other. An absolute value of "1" means the image is better. A "0" means the images are similar.

8 Results

The similarity scores are presented first. Table 2 provides the comparison between the FSL segmentations and the classification algorithms for all images. Results indicate *k*-NN performs slightly worse than MLR. The spline model seems to pick up CSF better than the linear MLR model, and the linear MLR model classifies WM better than the spline MLR model.

The confusion matrices for the FSL segmentations are presented in Table 5. These matrices provide the classifications and misclassifications for all subjects combined. The diagonal elements represent the probability of a voxel being classified as tissue j given that the gold standard classifies it as tissue j. No one method is superior, but *k*-NN does have somewhat smaller diagonal entries than the other two methods. Confusion matrices show that the spline MLR model predicts CSF better and the linear MLR model predicts WM better. For all methods the probability of a voxel being classified as WM when the gold standard labels it as CSF, and vice versa, is 1% or less.

Figure 5 shows the results of the classifications and compares them to their FSL segmentations. It is difficult to compare the different models and methods based on these images because they appear similar. The difference images show a fairly dense set of voxels that are misclassified, especially at the brainstem. In the slices presented here, it appears that the spline MLR model has a higher misclassification rate than the linear MLR model.

Tables 3 and 6 provide the results with comparison to FreeSurfer segmentations. Here the similarity indices are low for CSF. An explanation is presented in the next section. All three statistical methods have roughly equivalent similarity with the FreeSurfer segmentations.

Tables 4 and 7 display the similarity indices between the results of the proposed method and TOADS segmentation software. The results from *k*-NN appear to be similar to the MLR models in WM only. For the other two tissues, the MLR models appear to match TOADS better. The confusion matrices in Table 7 tell the same story.

The segmentations from the software and the classification methods were compared and scored by a radiology resident, and the results are presented in Figure 5. All three classification algorithms are preferred over TOADS and FreeSurfer, where the mean score is approximately 1 for all algorithms. Compared to FSL, both *k*-NN and the full MLR model have a mean slightly less than 0, indicating that FSL is slightly preferred to these models. However, the spline MLR model has a mean slightly larger than 0, indicating that it is preferred to FSL. All three means are approximately 0, from which it is concluded that FSL segmentations are preferred approximately equally to the model classifications.

One important result is captured in Figure 7. Partial ROC curves are presented for the three different tissue classes, comparing the spline MLR model to a more traditional MLR model that is based on the four most common images from multi-sequence analyses: T1, T2, FLAIR, and PD. In the figure, this is labeled as the "primitive" MLR model. The partial ROC curves (with a false positive rate bounded between 0 and 0.1) for the spline MLR model have a larger AUC in eight of the twelve comparisons. The ratios of the partial AUC for the spline MLR model to the corresponding partial AUC of the primitive MLR model has mean 1.007 and standard deviation 0.032. The full ROC curves for the spline model have a larger AUC in eleven of the twelve comparisons. The ratios of the full AUCs between the spline MLR and the primitive MLR models have mean 1.003 with standard deviation 0.005. Thus, while the spline MLR model performs better, using the primitive MLR model performed almost as well. From a practical perspective, the method proposed

here can be easily applied to most studies that contain T1, T2, FLAIR, and PD images. The loss of prediction power is quite minimal.

9 Discussion

Methods for local segmentation are presented and performed in this study and compared to the FSL, FreeSurfer, and TOADS-CRUISE software. Local segmentation uses only a set of covariates specific to a voxel—what happens at one voxel has no effect on neighboring voxels. Interestingly, the proposed MLR method, using only local information, gives similar results to the automated software FSL. FSL takes a single T1 image as input, but it also incorporates spatial information through a hidden Markov random field. Compared to the other two software segmentations, FreeSurfer and TOADS, both atlas-based automated segmentation software, the proposed methods and their local segmentation produces results that are preferred.

Classifiers based on the new statistical methods and existing software have both advantages and disadvantages. With the FSL software, no training or initialization is required. The result is that each FSL segmentation starts from scratch, and the program converges to a result without the need for training data. The MLR statistical model requires initial training, in this case on three subjects, and the segmentation is as fast as evaluating a function of regression coefficients. An advantage of MLR models is that they require training only once and the model parameters can then be used for any segmentation algorithm, as long as data are pre-processed and intensity normalized using the same steps described in this document. FSL and other software typically only require a single T1 image. The newly proposed statistical methods can work with one or multiple images. In particular, we have used four for the primitive model, six for the spline MLR model, and twelve for the linear MLR model. The acquisition of 12 images at the MRI scanner is not difficult, but it does require the subject to stay longer in the MRI machine.

There are even more advantages to using MLR models. They are flexible and can incorporate a different number of image sequences. Furthermore, they can be adapted by changing the modeling using observed features of the data. Both training and deployment of the model is fast. See Table 8 for a summary of computational time. Information around the voxel, also known as a texture, can be incorporated easily. Examples are the skew and variance of intensities in a neighborhood of a voxel. Finally, MLR models provide an excellent platform for comparing the relative influence of various predictors on the prediction performance of the algorithms.

The visual differences between the newly proposed methods and the FSL segmentations are small. Even though the difference images show a non-negligible divergence in classification, the resulting segmentation images are probably not distinct enough for a human observer. This suggests that in the future it may be a good idea to provide a disagreement map to the human observer during visual inspection of the images. Even though the newly proposed statistical methods do not incorporate any spatial dependence, most areas of the brain are sufficiently smooth and their classification tends to be spatially consistent. One notable exception is in the brainstem. This indicates that a diverse enough set of images may lead to better segmentation.

The similarity results between MLR models and k-NN based on FreeSurfer tissue masks are not as good. There are at least two reasons for this. First, reclassification is not precise. For example, some anatomical features, such as the thalamus and pallidum, are a heterogeneous mixture of GM and WM, but the FreeSurfer masks labeled the entire region as GM. This causes obvious training errors. Moreover, the number of voxels labeled as CSF is two orders of magnitude smaller than the number of voxels labeled as GM or WM. The

high relative prevalence of GM and WM influences the poor similarities between the predicted classifications and the FreeSurfer masks.

The ratings from the radiology resident indicate that the proposed methods produce results that are as good or better than the ones obtained from automatic segmentation software. See Figure 6. For the subplot corresponding to FSL, each statistical method's distribution has mean near "0." This means that on average, the statistical models trained on FSL produce results similar to the segmentations from FSL. Furthermore, most ratings in the FSL subplot are between -1 and 1 for all statistical prediction algorithms with an even mix between -1, 0, and 1. This confirms that segmentations from FSL are not too different from segmentations from the three algorithms (*k*-NN, linear MLR, and spline MLR). In the subplots corresponding to TOADS and FreeSurfer, each statistical method's distribution is centered close to "1." This indicates that on average, each statistical method's results are better than the segmentations from TOADS and FreeSurfer.

By their nature, segmentations from both TOADS and FreeSurfer aggregate different anatomical features into CSF, GM, and WM. This may create conspicuous problems as some labeled structures contain a mixture of tissue classes. In contrast, FSL uses morphologic operations, and its segmentations compare well with the results from the new statistical algorithms. These differences may partly explain why the new statistical algorithms compare differently to the three automated software.

There are several limitations on this study. First, the number of subjects is very small. A larger number of healthy subjects would have permitted a more thorough analysis and possibly better generalization potential. However, a larger number of subjects may lead to serious computational challenges. In this study, the only computational challenge was raised by the *k*-NN classifier; the problem was resolved by aggressive sub-sampling. Second, the

lack of true gold standards means that while it is possible to compare on a relative scale one method to another, it is impossible to measure on an absolute scale the quality of any given method. Third, alignment, intensity inhomogeneity, and brain extraction induce some additional errors in the data. Better pre-processing pipelines may actually lead to large improvements in tissue-class classification.

There are many approaches that could be used to improve the performance of the methods introduced in this study, while additional segmentation algorithms could be used for comparison (e.g. SPM, FIL Methods group 2014). Also, the statistical models could be developed further to include additional covariates. One interesting kind of covariate is a statistic in a small neighborhood of a given voxel (e.g. standard deviation or skewness of the neighborhood voxel intensities), sometimes called a texutre. Furthermore, a spectrum of spatial information could be incorporated to evaluate the gain in precision compared to the increase in spatial information. The current study does not use spatial information.

Another way to improve the results of the statistical models is to perform some postprocessing. This could be a simple smoothing of the resultant probability maps or the hard classification results from MLR or *k*-NN. Another alternative is to use the results from the statistical models as the initialization for another volumetric, topology-based algorithm. Both of these approaches would be a way to classify a given voxel while incorporating information from nearby voxels.

With more MRI sequences being created, it only makes sense to include this information in statistical analyses. This study indicates that adding additional images to a feature space has the potential to improve results. Lastly, we have shown that even using the subset of the images that is typically acquired in most brain imaging studies provides prediction performance that is very close to that of models using all images. Being able to

quantify how close the prediction performance of two models is provides a lot of information about the need, or lack thereof, for additional image sequences.

10 Figures and Tables

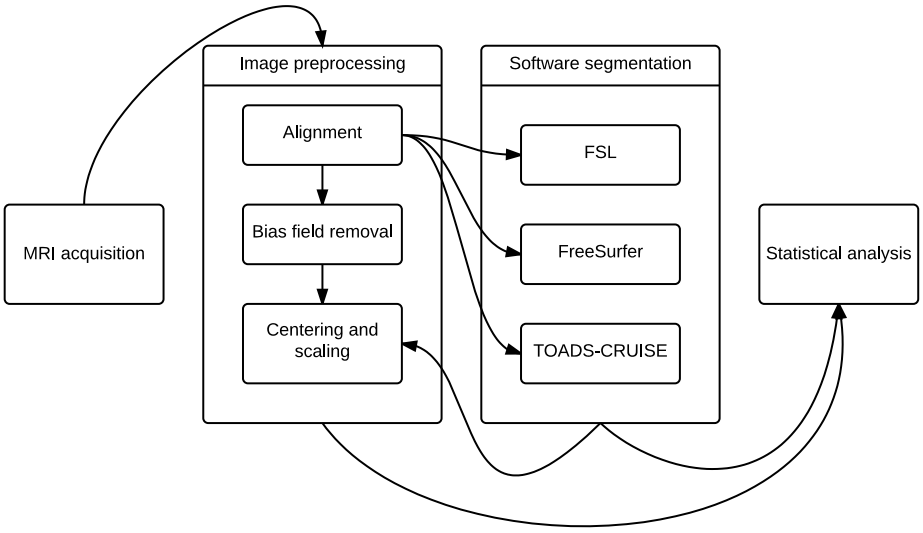

Figure 1: Pipeline for image preprocessing and generating software segmentations. After image acquisition, the water image serves as the master image for alignment. These aligned images serve as input to the software segmentations. Back in the image preprocessing, next the bias field removal is performed. This simultaneously corrects for intensity inhomogeneity and normalizes the image. Finally, the data are centered and scaled with respect to brain masks created by the software. Once these steps are completed, the data are ready for statistical analysis.

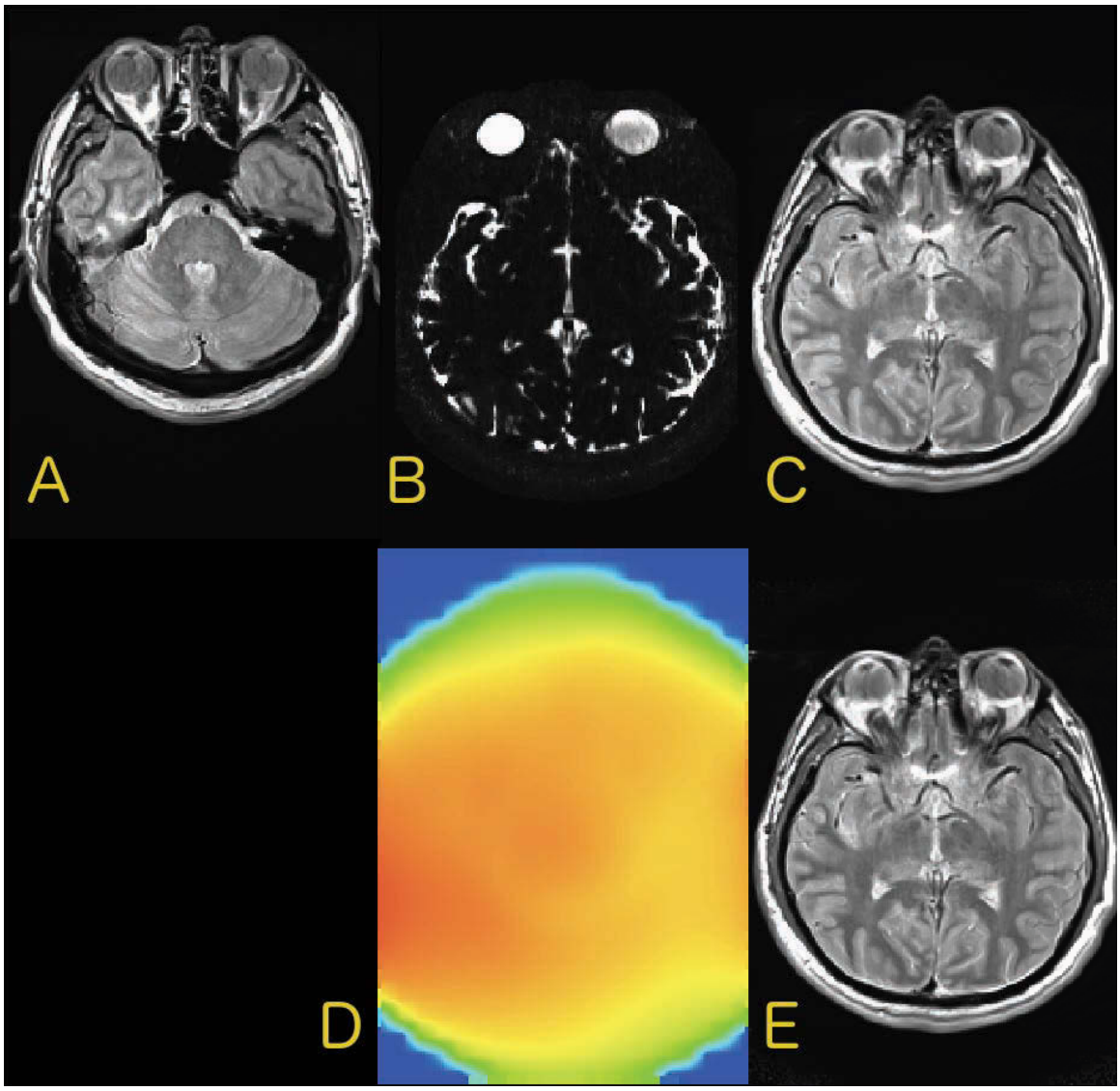

Figure 2: The path of an axial slice through the preprocessing pipeline. (A) shows the original PD image out of alignment with (B), the water image. (C) displays the PD image after alignment. An estimate of the bias field from (C) is presented in (D) as a heat map. Notice the left side has more red than the right, which is primarily yellow and orange. This indicates that the aligned image in (C) is brighter on the left side than on the right. Finally, (E) shows the PD image with the bias field corrected.

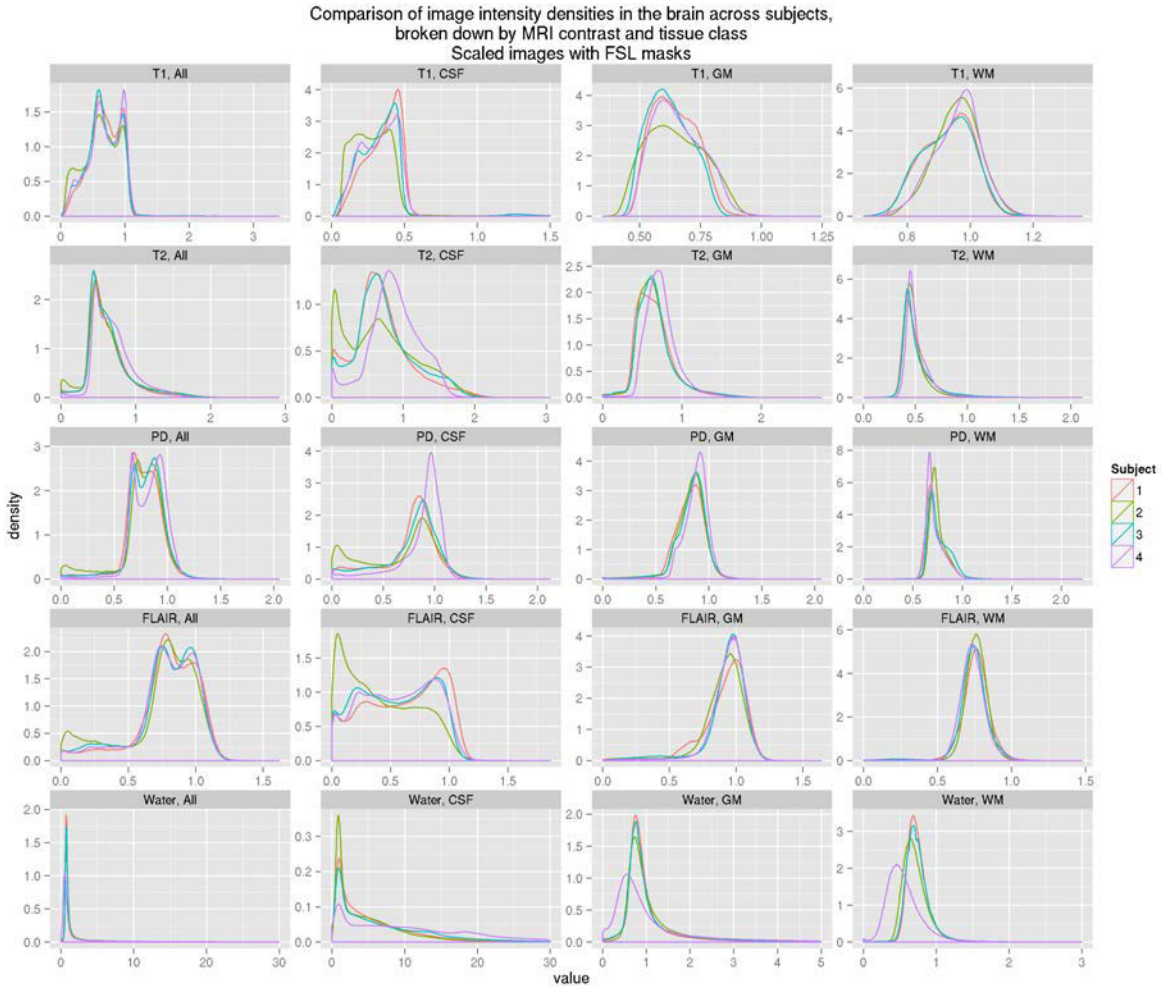

Figure 3: The densities of image intensities under various masks, compared across subjects. In each column, the densities of a specific mask are displayed. "All" stands for the entire brain. Each row pertains to a different image. For example, row two, column three shows the densities of the intensities in the GM mask on the T2 image for all four subjects.

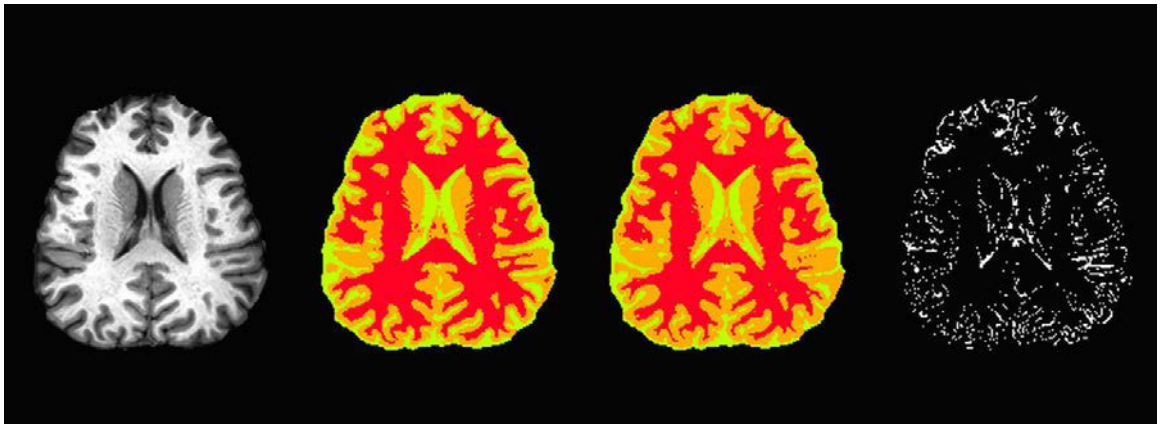

Figure 4: An example comparison between two segmentations (in the middle). The rater has to choose if one segmentation is better than the other. The corresponding T1 slice is on the left and the difference indicator (white only where the two segmentations differ) is on the right.

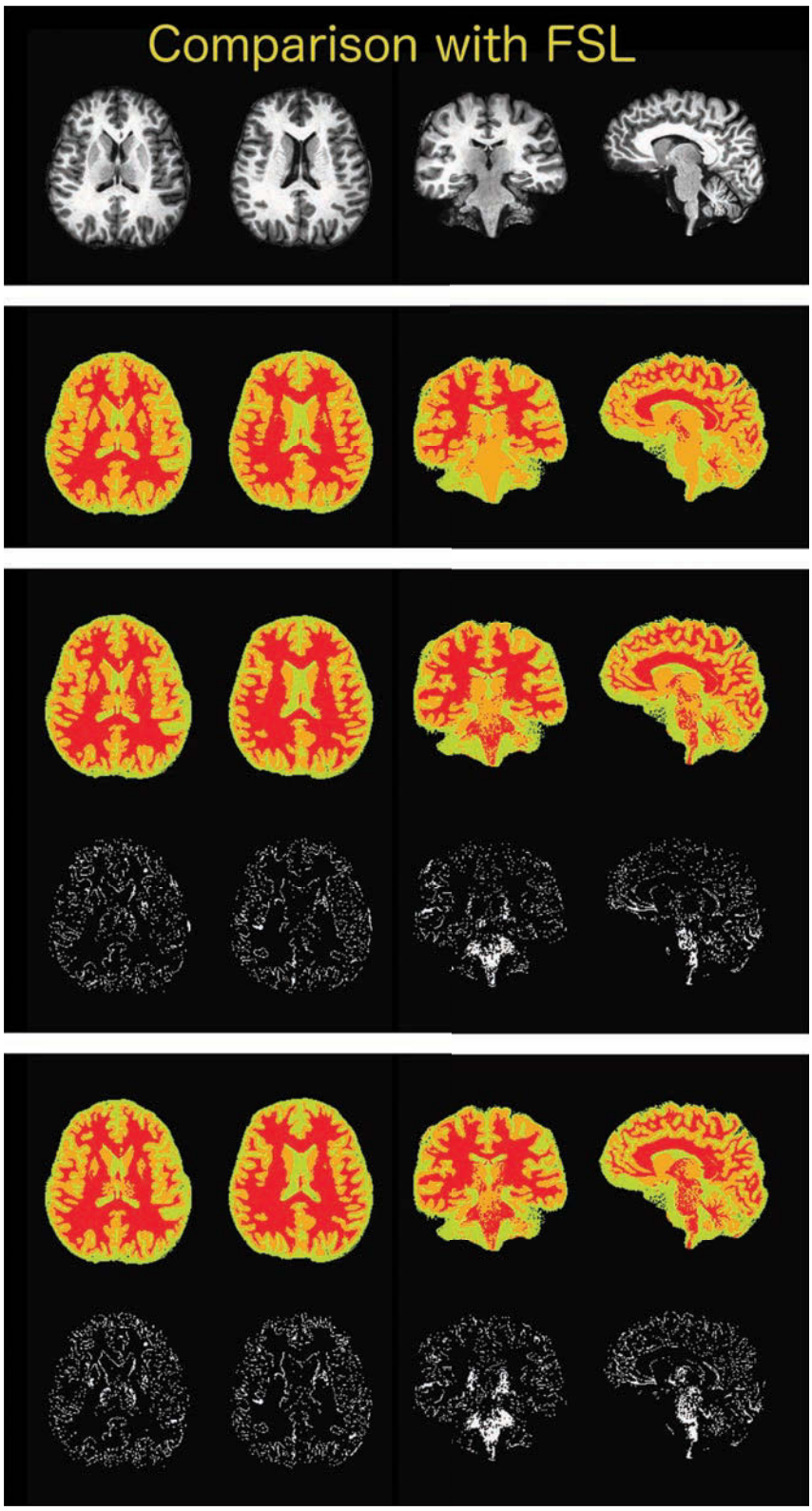

Figure 5 Slices of interest and various segmentation results. Each column has information about a common slice. The top row has the original T1 image. The second row shows the FSL segmentation. The third shows the output of the full MLR model, and the fourth row shows the differences between that classification and the FSL classification. The fifth and sixth rows show the results for the spline MLR model.

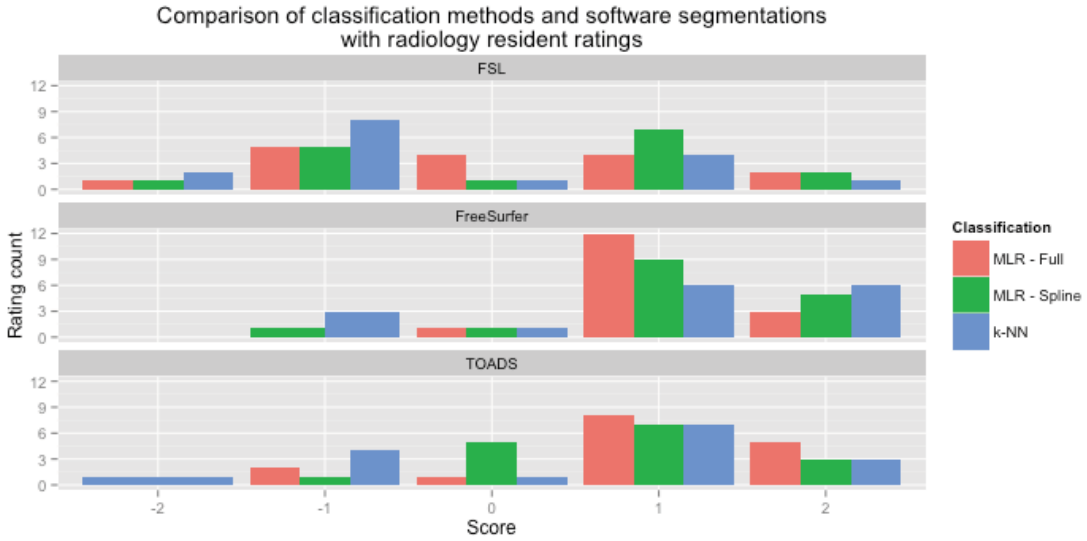

Figure 6: Ratings of classifications compared to software segmentations. Each classification method is compared with each software segmentation on the same predetermined four slices per subject. The bars show how many of each score was chosen. The scores mean the following: -2 indicates the software segmentation is much better; -1 indicates the software segmentation is better; 0 indicates the two are similar; 1 indicates the given classification method is better; 2 indicates the given classification is much better. The images were rated by a radiology resident.

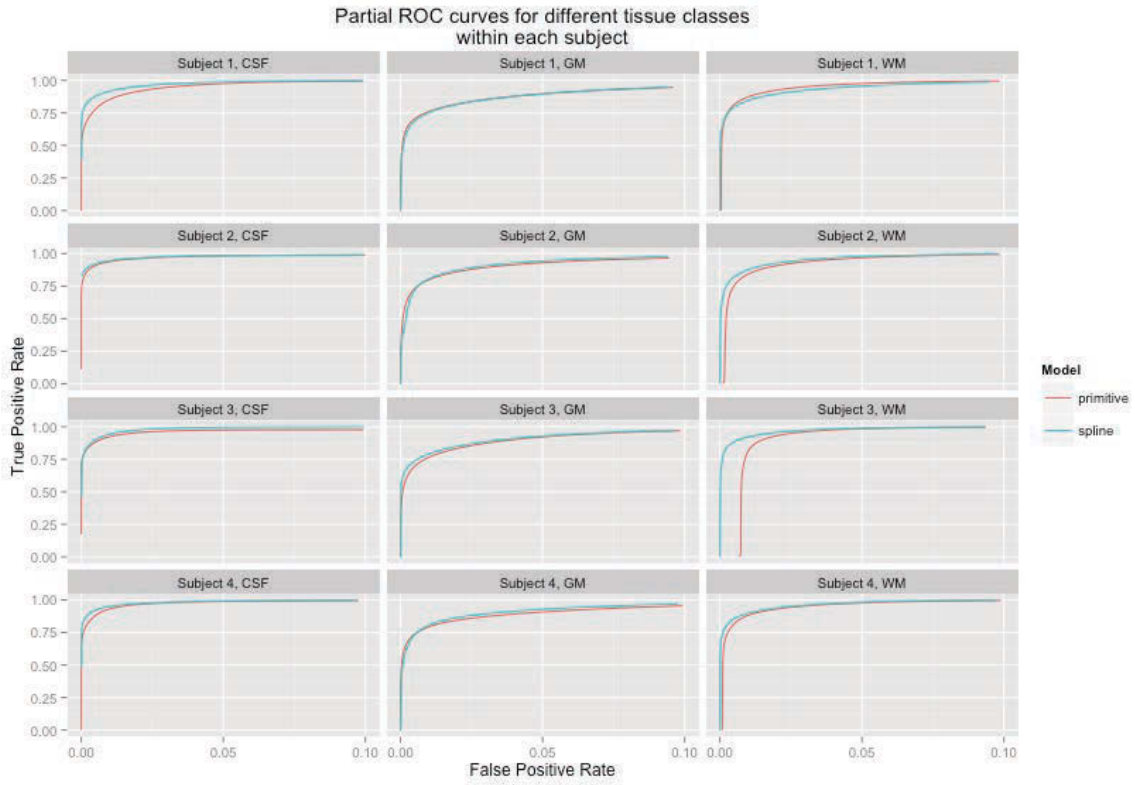

Figure 7: ROC curves for each subject and tissue class. Both primitive and spline MLR models produce probabilities that a voxel belongs to a certain tissue class. These were referenced against the FSL segmentations. In eight of the twelve cases, the spline model has a larger partial AUC.

	Type	FA (degrees)	TR (ms)	TE (ms)	TI (ms)	Resolution (mm)
T1	GRE	18	8.1	3		1x1x1
T2	FSE		5000	18		0.94x0.94x3
PD	FSE		5000	82		0.94x0.94x3
FLAIR	IRFSE		4800	350	1800	1x1x1
Water	FSE		4800	750		0.67x0.67x0.67
MTC	GRE		35	2		1x1x1

Table 1: A description of the MRI sequence protocols. The acronyms and their meanings are as follows: FA is flip angle; TR is repetition time; TE is echo time; TI is inversion time; GRE is gradient recalled echo; FSE is fast spin echo; IRFSE is inversion recovery fast spin echo. For the resolution, the third dimension is the slice thickness. All images obtained in this study use this protocol.

		FSL					
		Jaccard index			Dice index		
Model	Subject	CSF	GM	WM	CSF	GM	WM
MLR	1	0.740	0.818	0.887	0.851	0.900	0.940
Model 1	2	0.892	0.831	0.809	0.943	0.908	0.895
Full	3	0.892	0.856	0.831	0.943	0.922	0.908
	4	0.877	0.868	0.903	0.934	0.929	0.949
	All	0.856	0.843	0.859	0.923	0.915	0.924
MLR	1	0.821	0.824	0.844	0.901	0.903	0.915
Model 2	2	0.867	0.810	0.809	0.927	0.895	0.894
Spline	3	0.915	0.865	0.846	0.956	0.927	0.917
	4	0.890	0.873	0.903	0.942	0.932	0.949
	All	0.875	0.843	0.852	0.933	0.915	0.920
k-NN	1	0.748	0.766	0.802	0.856	0.867	0.891
	2	0.878	0.818	0.799	0.935	0.900	0.889
	3	0.806	0.822	0.839	0.893	0.903	0.912
	4	0.849	0.855	0.909	0.918	0.922	0.952
	All	0.825	0.815	0.839	0.904	0.898	0.912

Table 2: The similarity results comparing the statistical and algorithmic output with the FSL segmentation. There is a comparison for each tissue class in each subject.

		FreeSurfer					
		Jaccard index			Dice index		
Model	Subject	CSF	GM	WM	CSF	GM	WM
MLR	1	0.499	0.895	0.865	0.666	0.944	0.928
Model 1	2	0.549	0.889	0.860	0.709	0.941	0.925
Full	3	0.642	0.844	0.790	0.782	0.915	0.882
	4	0.759	0.855	0.830	0.863	0.922	0.907
	All	0.638	0.870	0.835	0.779	0.931	0.910
MLR	1	0.492	0.896	0.868	0.659	0.945	0.929
Model 2	2	0.573	0.886	0.854	0.729	0.940	0.921
Spline	3	0.605	0.849	0.799	0.754	0.918	0.888
	4	0.712	0.857	0.840	0.832	0.923	0.913
	All	0.613	0.871	0.840	0.761	0.931	0.913
k-NN	1	0.445	0.885	0.845	0.616	0.939	0.916
	2	0.471	0.883	0.852	0.640	0.938	0.920
	3	0.612	0.850	0.799	0.759	0.919	0.889
	4	0.744	0.872	0.853	0.853	0.932	0.921
	All	0.602	0.872	0.837	0.751	0.932	0.911

Table 3: The similarity results comparing the statistical and algorithmic output with the FreeSurfer segmentation.

		TOADS					
		Jaccard index			Dice index		
Model	Subject	CSF	GM	WM	CSF	GM	WM
MLR	1	0.807	0.854	0.868	0.893	0.921	0.930
Model 1	2	0.867	0.860	0.859	0.929	0.925	0.924
Full	3	0.844	0.872	0.876	0.915	0.931	0.934
	4	0.777	0.785	0.835	0.875	0.879	0.910
	All	0.815	0.842	0.860	0.898	0.914	0.924
MLR	1	0.829	0.852	0.853	0.907	0.920	0.920
Model 2	2	0.879	0.855	0.849	0.935	0.922	0.918
Spline	3	0.866	0.873	0.873	0.928	0.932	0.932
	4	0.777	0.769	0.818	0.874	0.870	0.900
	All	0.826	0.837	0.848	0.905	0.911	0.918
k-NN	1	0.762	0.820	0.841	0.865	0.901	0.914
	2	0.804	0.842	0.857	0.891	0.914	0.923
	3	0.786	0.853	0.875	0.880	0.921	0.934
	4	0.746	0.747	0.823	0.855	0.855	0.903
	All	0.769	0.815	0.849	0.869	0.898	0.918

Table 4: The similarity results comparing the statistical and algorithmic output with the TOADS segmentation.

		MLR Model 1 – Full			
		CSF	GM	WM	n
FSL	CSF	0.901	0.091	0.008	1462354
	GM	0.027	0.927	0.046	2843272
	WM	0.000	0.078	0.922	1933222
		MLR Model 2 – Spline			
		CSF	GM	WM	n
FSL	CSF	0.923	0.076	0.001	1462354
	GM	0.028	0.924	0.048	2843272
	WM	0.001	0.085	0.914	1933222
		k-NN			
		CSF	GM	WM	n
FSL	CSF	0.875	0.119	0.007	1462354
	GM	0.031	0.912	0.057	2843272
	WM	0.000	0.087	0.913	1933222

Table 5: Confusion matrix for the FSL segmentation and the statistical and algorithmic output. For each row, the numbers represent the proportion of that row's FSL tissue class being classified as the tissue class in the column header from the results of the model above the sub-table. The column headed as "n" shows the number of voxels in that row's FSL tissue class.

		MLR Model 1 – Full			
		CSF	GM	WM	n
FreeSurfer	CSF	0.700	0.293	0.007	85634
	GM	0.003	0.936	0.061	2715353
	WM	0.001	0.092	0.907	1947183
		MLR Model 2 – Spline			
		CSF	GM	WM	n
FreeSurfer	CSF	0.686	0.307	0.007	85634
	GM	0.003	0.934	0.063	2715353
	WM	0.001	0.086	0.914	1947183
		k-NN			
		CSF	GM	WM	n
FreeSurfer	CSF	0.656	0.340	0.004	85634
	GM	0.002	0.942	0.056	2715353
	WM	0.001	0.097	0.902	1947183

Table 6: Confusion matrix for the FreeSurfer segmentation and the statistical and algorithmic output.

		MLR Model 1 – Full			
		CSF	GM	WM	n
TOADS	CSF	0.890	0.102	0.007	1091962
	GM	0.033	0.921	0.046	3092175
	WM	0.000	0.081	0.919	2200584
		MLR Model 2 – Spline			
		CSF	GM	WM	n
TOADS	CSF	0.900	0.092	0.008	1091962
	GM	0.032	0.914	0.055	3092175
	WM	0.000	0.083	0.916	2200584
		k-NN			
		CSF	GM	WM	n
TOADS	CSF	0.872	0.120	0.007	1091962
	GM	0.047	0.897	0.056	3092175
	WM	0.000	0.081	0.918	2200584

Table 7: Confusion matrix for the TOADS segmentation and the statistical and algorithmic output.

Method	Mean time	Standard deviation
Linear MLR – training	12.18 min	4.45 min
Linear MLR – prediction	13.33 sec	3.42 sec
Spline MLR – training	27.21 min	8.76 min
Spline MLR – prediction	16.5 sec	5.63 sec
k-NN	45.44 min	6.34 min

Table 8: Computational times to run full segmentation algorithms. For the MLR models, the mean and standard deviation are calculated for the times to train on three subjects and to test on one subject. For the *k*-NN model, the mean and standard deviation are calculated of the overall times to train on 5% of the voxels from three subjects and test on one subject. Calculations were performed on the Joint High Performance Computing Exchange (JHPCE) at Johns Hopkins University.

11 Bibliography

Agresti, Alan. *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc, 2002.

Cotten, Anne, Erwan Kermarrec, Antoine Moraux, and Jean-François Budzik. "New MRI Sequences." *Joint Bone Spine* 76 (2009): 588-590.

Devroye, Luc, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* New York: Springer, 1996.

Fischl, Bruce, et al. "Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain." *Neuron* 33 (2002): 341-355.

González Ballester, Miguel Ángel, Andrew Zisserman, and Michael Brady. "Segmentation and measurement of brain structures in MRI including confidence bounds." *Medical Image Analysis* 4 (2000): 189-200.

Gordillo, Nelly, Eduard Montseny, and Pilar Sobrevilla. "State of the art survey on MRI brain tumor segmentation." *Magnetic Resonance Imaging* 31 (2013): 1426-1438.

Hutton, Chloe, Enrico De Vita, John Ashburner, Ralf Deichmann, and Robert Turner. "Voxel-based coritcal thickness measurements in MRI." *NeuroImage* 40, no. 4 (May 2008): 1701-1710.

Kowkabzadeh, Koushyar. "Evaluations of Tissue Segmentation of brain MR Images." Gothenburg: Chalmers University of Technology, 2010.

Rivest-Hénault, David, and Mohamed Cheriet. "Unsupervised MRI segmentation of brain tissues using a local linear model and level set." *Magnetic Resonance Imaging* 29 (2011): 243-259.

Shinohara, R. T., C. M. Crainiceanu, B. S. Caffo, M. I. Gaitan, and D. S. Reich. "Population-wide principal component-based quantification of blood-brain-barrier dynamics in multiple sclerosis ." *NeuroImage* 57, no. 4 (2011): 1430-1446.

Shinohara, R. T., et al. "Statistical normalization techniques for magnetic resonance imaging." *NeuroImage: Clinical* 6 (2014): 9-19.

Smith, Stephen M., et al. "Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis." *NeuroImage* 17, no. 1 (2002): 479-489.

Sweeney, E. M., R. T. Shinohara, C. D. Shea, D. S. Reich, and C. M. Crainiceanu. "Automatic lesion incidence estimation and detection using multisequence longitudinal MRIs." *American Journal of Neuroradiology* 34, no. 1 (2013): 68-73.

Sweeney, Elizabeth M., et al. "OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI." *NeuroImage: Clinical* 2 (2013): 402-413.

Tustison, N. J., et al. "N4ITK: improved N3 bias correction." *IEEE Trans Med Imaging* 29, no. 6 (Jun 2010): 1310-1320.

Tustison, Nicholas J., et al. "Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements." *NeuroImage* 99 (2014): 166-179.

Vovk, Andrej, Robert W. Cox, Janez Stare, Dusan Suput, and Ziad S. Saad. "Segmentation priors from local image properties: Without using bias field correction, location-based templates, or registration." *NeuroImage* 55 (2011): 142-152.

Withey, Daniel J., and Zoltan J. Koles. "A Review of Medical Image Segmentation: Methods and Available Software." *International Journal of Bioelectromagnetism* 10, no. 3 (2008): 125-148.

12 Scholarly life

James Kenneth Pringle was born on 20 June 1988, in Dayton, Ohio, to Darl and Maria Pringle. He attended college at Brigham Young University and graduated *magna cum laude* and with University Honors in April 2012. He studied Mathematics, Russian, Physics, and Computer Science.

Following college graduation, James attended graduate school in the Biostatistics Department of Johns Hopkins University Bloomberg School of Public Health. He participated in summer research projects in 2013 and 2014, the first with Johns Hopkins Oncology Biostatistics, and the second with NINDS at the National Institutes of Health. His studies and research at the NIH became the foundation for his Master's thesis. He graduated with the ScM degree in December 2014.